Review of Influencing Factors and Prediction of Individual Paper Citation Count: Postprint

Authors: Zhang Sufang, Liu Huimin

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] To review the influencing factors of citation frequency for individual papers and the current state of citation frequency prediction research, thereby providing researchers and research institutions with a comprehensive and systematic cognitive framework for investigating the influencing factors and prediction of citation frequency for individual papers. [Process/Method] Employing the literature survey method, this study systematically reviews existing literature to summarize the content and characteristics of influencing factors, research objects, and research methods related to citation frequency prediction. Through tabular comparison, different approaches are analyzed, and both common problems in existing research and some innovative solutions are identified. [Results/Conclusion] The systematic review and summary reveal that the causal relationship between influencing factors and prediction results is unclear, research sample data lacks diversity, the relationship between the applicability of research findings and prediction periods is not explicitly defined, and the interpretability of model evaluation is weak. Therefore, future research quality should be enhanced by addressing prerequisites for problem-solving, selecting targeted samples, improving methods for extracting influencing factors, and employing mathematical thinking in modeling.

Full Text

A Review of Research on Influencing Factors and Prediction of Citation Frequency of a Single Paper

Zhang Sufang, Liu Huimin
School of Economics and Management, South China Normal University, Guangzhou 511400

Abstract:
[Purpose/Significance] This paper systematically reviews the relevant influ-

encing factors of single-paper citation frequency and the current state of citation frequency prediction research, providing a comprehensive cognitive framework for researchers and institutions studying these topics. [**Method/Process**] Using the literature research method, we systematically combed through existing literature to summarize the influencing factors, research objects, and research methods of citation prediction, compared different approaches through tabular analysis, and identified common problems and innovative solutions in current research. [**Result/Conclusion**] Our systematic review reveals that the causal relationship between influencing factors and prediction results is unclear, research sample data lacks diversity, the relationship between result applicability and prediction period is not explicitly defined, and model evaluation interpretability is weak. Therefore, future research quality should be improved by addressing problem preconditions, selecting targeted samples, refining influencing factor extraction methods, and employing mathematical thinking in modeling.

Keywords: citation frequency prediction; influencing factors; regression analysis; machine learning; deep learning

The scientific system contains numerous elements and connections, and researchers are increasingly interested in the citation dynamics of academic papers and scientific evolution. Citation frequency reflects the attention a paper receives to some extent; however, typically only a small minority of papers accumulate the vast majority of citations, while most others attract only a few [**?**]. In other words, some research papers are more likely than others to attract researchers' attention. Given the growing volume of literature, predicting which papers will attract academic attention is important. Consequently, citation frequency prediction has emerged as a new research direction in bibliometrics. Numerous papers have appeared on this topic, with researchers often troubled by large numbers of low-citation papers during modeling. The diverse selection of methods and influencing factor features has led to redundant research. Although scholars have conducted systematic reviews of this topic, they have focused primarily on influencing factors and research methods, without proposing effective solutions for how researchers can engage with this field. Therefore, this paper reviews the influencing factors of paper citation frequency, addressing prediction tasks by examining research methods, object forms, and prediction periods for single-paper citation frequency from both regression and classification perspectives, and finally proposes solutions to common problems in existing research to provide reference for future studies. The main framework of this review is shown in Figure 1 [Figure 1: see original paper].

Influencing Factors of Single-Paper Citation Frequency

Citation frequency prediction for academic papers has been extensively studied. Researchers typically focus on what factors affect citation counts to screen important influencing factors for prediction. F. Didegah and M. Thelwall [**?**] argue that citation motivations are complex; the intellectual cognition of papers by citers constitutes an intrinsic factor that can be investigated through

interviews and questionnaires, but this approach is time-consuming and, due to the complexity and discipline-dependency of citation motivations, such qualitative research usually involves only a small sample of scholars. External factors, however, can be quantified and calculated on a large scale and thus used to predict future citation impact. External factors affecting citation rates include attributes of the cited paper such as authors, abstracts, journals, fields, references, and the paper itself. This study is limited to external motivations, categorizing these factors into four major types: paper-related, author-related, journal-related, and other factors.

Paper-Related Influencing Factors

Among paper-related influencing factors, one of the main factors associated with citation frequency is the paper' s topic, which represents the core of its research content and can be used to predict future citation frequency [**?**]. Content can be analyzed from three dimensions: topic attention (the attention the paper' s topic receives), topic novelty, and topic diversity. Popular topics typically attract more attention and citations from other papers [**?**], while novel topics can enhance impact and citation rates [**?**]. The more attractive and novel the topic, the more citations the paper may receive. Additionally, the scope and field of the research topic affect citation frequency, as topic diversity influences citation counts [**?**].

In topic identification research, most researchers use the Latent Dirichlet Allocation (LDA) model or its derivatives for topic identification, then calculate metrics such as topic attention/heat, novelty, and diversity. Topic attention is primarily measured from the perspective of cumulative citations, diversity from information entropy, and novelty from peer review, citation, and content perspectives [**?**]. Apart from peer review, the other two methods (co-occurrence frequency of citation pairs and topic content) are based on a co-occurrence concept. Numerous studies exist on content novelty, though their perspectives are largely similar.

The number, authority, and diversity of references also increase citation frequency [**?**]. Studies with more references correlate with higher citation rates [**?**]. Papers with younger average reference ages may receive more citations, while those citing "old publications" show significantly reduced citation counts [**?**], as paper information becomes outdated over time [**?**]. Generally, citation frequency peaks in the first few years after publication and gradually decreases over time. Additionally, reference authority (cumulative citation frequency [**?**]) and diversity (citing literature' s research fields [**?**] and cross-nationality [**?**]) affect citation rates.

Research also finds that certain document types receive more citations than others; for example, review papers are cited more than research papers [**?**]. Funding is an important economic source for scientific research, and adequate funding provides better material support. Generally, papers from higher-level

funded projects receive more citations than unfunded papers [?]. In some studies, early citation rates and their velocity are considered predictors of future citations [?]. Early citations represent the scientific community' s initial feedback on a paper, and citation velocity reflects the paper' s dissemination speed within the scientific community. Paper length (number of pages) also increases citation frequency [?], as longer papers contain more information [?]. The title is the most condensed summary of a paper' s content and the first thing scholars see when searching. H. R. Jamali and M. Nikzad [?] found that an informative title can increase citations, but no significant correlation exists between title length and citations; title features affect download counts more than citation frequency [?]. Open access refers to paper accessibility and visibility—authors must be able to read the full text to benefit from it—and open access papers typically receive more citations than those in non-open access journals [?][?].

The scope of paper-related influencing factor research is extensive. Beyond the well-studied factors mentioned above, foreign scholars have conducted detailed research on methodology/research design, section characteristics, and data/appendix usage [?]. Although some papers find correlations between these factors and citation rates in certain fields, these factors may not relate or only weakly relate to citation rates in different fields. These studies often ignore disciplinary differences [?], yet some influencing factors have obvious disciplinary characteristics. Therefore, building universal comprehensive indicators is not an ideal choice. The above influencing factors are summarized in Table 1 .

Author-Related Influencing Factors

Author-related factors also affect citation frequency. Author count is a measure of research collaboration. High-quality papers often involve cooperation among multiple researchers, and co-authorship (especially international collaboration [?]) can increase citation rates [?]. However, some studies have found opposite results, showing no special connection between international cooperation and citation frequency [?]. As the citation time window lengthens, the correlation between author count and citations weakens [?]. Yet other reports indicate that author collaboration in different fields can increase citation rates [?]. Therefore, whether collaboration affects citation frequency remains controversial.

Additionally, author count correlates positively with self-citation numbers [?], but the ratio of self-citation to non-self-citation decreases as total accumulated citations increase, with self-citations often concentrated shortly after publication [?]. Thus, from a macro perspective, self-citations need not be excluded when analyzing paper citations [?].

Famous authors have high prestige in their fields, and their papers often receive more citations [?]. The Matthew Effect makes papers by high-citation authors more likely to be cited than those by low-citation authors [?]. Therefore, an author' s previous papers' citation frequency can be a good predictor of future citations [?]. The H-index is the most common standard for measur-

ing researcher capability [**?**], and prestigious authors often have high H-indices. Thus, when studying how an author' s prestige in a field affects citations, the H-index is commonly used as a metric. Institutional prestige largely depends on the author; generally, papers from top-ranked universities receive more citations [**?**].

Beyond these factors, demographic characteristics of authors have been included in measurement indicators. Some studies find that white males have higher influence than non-whites and females [**?**], while others show demographic characteristics have no significant effect on whether a paper is cited [**?**]. Author-related influencing factors are summarized in Table 2 .

Journal-Related Influencing Factors

Beyond paper and author factors, research finds that journal-level factors are the main determinants of citation frequency [**?**]. The average citation count a paper receives in its publishing journal can predict future citations [**?**]. Researchers tend to publish in high-impact journals to increase visibility and obtain higher citations. Studies prove that publishing in high-impact journals yields higher citations than low-impact journals [**?**]. Although numerous studies confirm the positive correlation between journal impact and paper citations, some find that journal impact factor is not necessarily a predictive indicator [**?**]. Some researchers use total citations and productivity (number of published papers) as influencing factors [**?**]. Additionally, some studies suggest journal language affects citation rates [**?**], particularly English journals, which accumulate more citations [**?**]. Journal-related influencing factors are summarized in Table 3 .

Other Influencing Factors

As research deepens, new perspectives such as social networks and time factors have emerged. Researchers began analyzing potential connections between social network activity and bibliometrics [**?**]. Kong Ling et al. [**?**] added altmetric factors when summarizing relevant influencing factors, but altmetric factors target open academic networks and social media, differing from traditional academic paper websites. Beyond social networks, academic citation networks are also important. To measure author sociality, R. Yan et al. [**?**] built an author collaboration network and recursively calculated sociality using PageRank. Since academic paper citations have a half-life attribute, time factors are valuable for citation frequency prediction. E. Butun and M. Kaya [**?**] combined author citation networks with time factors, introducing a temporal link metric that considers the evolution of author citation networks, using local and global topological structures in complex networks to predict link weights based on links in citation networks—this was the first study to use directed, weighted, and temporal citation networks for citation frequency prediction. Other factors are summarized in Table 4 .

Methods for Predicting Single-Paper Citation Frequency

With the development of scientometrics, numerous research methods have been introduced into citation frequency prediction. From a task-oriented perspective, prediction can be defined as either a regression or classification problem. For regression, main methods include traditional regression analysis, machine learning, and deep learning. For classification, machine learning methods are primarily used. Each method has distinct characteristics.

Regression-Based Prediction Methods

Regression-based prediction uses a paper' s relevant features to predict its citation frequency at a certain time point [**?**]. Regression is currently the most commonly used prediction method [**?**]. This section reviews citation frequency prediction research from three aspects: traditional regression, machine learning, and deep learning.

Traditional Regression Prediction Early researchers often used traditional linear regression for fitting studies. C. Lokker et al. [**?**] used 17 reference-related features and 3 journal-related features to predict two-year citations for clinical papers, achieving a coefficient of determination (r^2) of 0.60 for the training set and 0.56 for the test set. In sensitivity analysis, specificity for papers in the top half and top third of citation rankings was 72% and 82%, respectively, showing that regression prediction works better for highly cited papers—a finding also reflected in G. Abramo et al. [**?**], which reflects the fact that most papers are low-cited while only a small portion are highly cited [**?**]. T. Yu et al. [**?**] used stepwise multiple regression to select good feature variables from paper external features, author features, journal features, and cited paper features to build a model describing the relationship between features and citation impact for predicting citations five years after publication. L. Bornmann et al. [**?**] used all papers published in 1980 from the WoS database across all disciplines (approximately 500,000 papers), using citation frequency in the 31st year after publication as the dependent variable for long-term impact prediction. They found that only citations in the first few years after publication significantly improved long-term impact prediction, a result also found by G. Abramo et al. G. Abramo et al. [**?**] used two linear regression models, finding that average prediction accuracy was good for citation windows longer than two years, with a three-year window sufficient for predicting long-term scientific impact. The model had low prediction accuracy for low-cited papers and varying accuracy across disciplines. Cheng Zixuan et al. [**?**] used stepwise regression to predict citations seven years after publication for library and information science journal papers, identifying 10 factors significantly correlated with citation frequency.

Traditional regression analysis is statistics-based. Such models are effective for small data volumes and simple relationships, with intuitive understanding and interpretation, but they have strict requirements for data distribution and lower

precision for complex data structures. Some papers using traditional regression for citation prediction are shown in Table 5 .

Machine Learning Prediction With technological development, machine learning has emerged in citation frequency prediction research. R. Yan et al. [**?**] used basic features of highly cited papers and multiple machine learning methods to predict citation frequency, with the best model (CART classification and regression tree) achieving an average r^2 of 0.786 for predicting 10-year citations. They found that author expertise and journal impact were significant influencing factors, while isolated content features could not predict citations. T. Chakraborty et al. [**?**] argued that most regression methods assume all published papers have similar citation patterns, which affects prediction accuracy. They proposed a stratified learning approach, dividing papers into six citation patterns and using support vector machines for regression on each pattern. Their study proved stratified learning effective but only for papers with average annual citations greater than 1. J. Chen and C. Zhang introduced IBM models to extract content features for calculating association probabilities between paper topics, used bipartite network projection to obtain author collaboration networks, and applied Gradient Boosting Regression Trees (GBRT) to predict citation counts. Experiments showed GBRT's "content feature"group performed best on the KDDCUP dataset [**?**]. However, X. Zhu and Z. Ban [**?**] used the ArnetMiner dataset with academic network features and found author features more important, with Support Vector Machines (SVM) achieving the highest r^2 of 88.87%. Some papers using machine learning for citation prediction are shown in Table 6 .

Deep Learning Prediction In recent years, deep learning methods such as neural networks have been applied to citation frequency prediction. Deep learning models are a special type of machine learning that allows models to learn data representations with multiple abstraction levels through multiple processing layers [**?**]. In deep learning, time series neural networks like RNN, LSTM, and GRU can predict future sequence values, while BP neural networks and CNN are more effective for feature processing.

A. Abrishami et al. [**?**] used RNN to learn paper citation sequences to predict future citation sequences, but only used early citation features without incorporating other information sources like author functions or paper content. LSTM is a variant of RNN. S. Yuan et al. [**?**] combined four phenomena—paper intrinsic quality, aging effect, Matthew effect, and recency effect—to propose citation frequency prediction models based on RNN and LSTM, but only used time series without author, journal, or paper features. In contrast, J. Wen et al. [**?**] extracted features for citation prediction and input them into GRU neural networks, comparing results with other regression models. Experiments showed high prediction accuracy and fast convergence, with citation time series prediction outperforming existing methods.

Unlike time series prediction methods, X. Ruan et al. [**?**] used a four-layer Back Propagation (BP) neural network to predict total future citations, finding BP neural network performance significantly superior to six baseline models (XGBoost, RF, LR, SVR, KNN, RNN). For prediction accuracy, low-cited papers had higher accuracy than high-cited papers. J. Xu et al. [**?**] proposed a data-centric approach combining many literature features to predict long-term scientific impact using Convolutional Neural Networks (CNN).

Unlike linear regression models that rely on statistics, deep learning methods have no strict requirements for experimental data distribution, and neural network predictions are typically robust. Additionally, shallow machine learning model performance depends on feature engineering quality—the better the feature engineering, the higher the model learning efficiency. However, feature engineering construction, selection, and extraction are not easy tasks. In contrast, deep neural networks have advantages in feature learning—automatic feature engineering [**?**]—which can automatically transform initial "low-level" feature representations into "high-level features" through multi-level and non-linear transformations [**?**]. Some papers using deep learning for citation prediction are shown in Table 7 .

Summary Most of the above prediction studies filter papers by removing low-cited papers before prediction because low-cited papers do not perform well in regression prediction, which is often only suitable for predicting highly cited papers. However, for newly published papers, we cannot know whether they are highly cited, creating a large gap between prediction effectiveness and practical application. Y. Dong et al. [**?**] argue that citation frequency prediction has a long-tail effect and is not suitable for regression because prediction effectiveness is fundamentally limited by the power-law distribution of citations—low-cited papers are common while highly cited papers are rare. Since the vast majority of papers accumulate few citations, traditional regression analysis struggles to measure citation frequency. To address this difficulty, extracting features of highly cited papers and mapping them to citation frequency can improve prediction efficiency to some extent. However, because low-cited papers are too numerous, features of highly cited papers are not obvious, greatly reducing prediction effectiveness on practical application datasets.

Classification-Based Prediction Methods

Transforming citation prediction from regression to classification makes prediction results more consistent with citation data distribution patterns, making models more generalizable despite coarser prediction granularity [**?**]. Compared to regression methods, classification-based approaches are relatively uniform, primarily using various machine learning methods. As classification tasks are supervised learning, they require setting a classification threshold to determine each paper' s label. Common classification methods for citation prediction include Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors

(KNN), Logistic Regression (LRC), Decision Trees, Gradient Boosting Decision Trees (GBRT), Bagging (BAG), Random Forest (RF), XGBoost, and AdaBoost.

A. Ibanez et al. [**?**] divided papers into three categories—few citations ($ $1), some citations (2-4), and many citations (>4)—using machine learning methods like Naive Bayes, Logistic Regression, Decision Trees, and KNN to predict citations from year one to four, with Logistic Regression and Naive Bayes achieving the highest accuracy. L. Fu and C. Aliferis [**?**] used SVM in biomedicine to predict whether a paper' s 10-year citations would exceed thresholds (20, 50, 100, 500), achieving AUC (Area Under Curve) of 0.857-0.918. M. Wang et al. [**?**] divided 219 astronomy and astrophysics papers into high, medium, and low groups, using a multi-classifier system of five decision tree classifiers to achieve high classification ability, showing that paper internal quality and external features (primarily author and journal reputation) help improve citation prediction. Y. Dong et al. [**?**] found that an author' s publication topics and publishing journals determine whether a paper will contribute to the author' s h-index, while topic popularity and co-author influence are unrelated to prediction targets. Their best model achieved over 87.5% accuracy in predicting whether a paper would contribute to its primary author' s h-index within five years. Geng Qian et al. [**?**] found through extensive experiments that GBDT, XGBoost, and Random Forest have strong prediction ability, with better relative performance for longer prediction periods.

Machine learning can handle both regression and classification problems. Among many studies, ensemble machine learning methods and support vector machines show good prediction performance. Compared to predicting regression values, machine learning performs better in classification. Although classification prediction is coarser-grained, it better fits practical application data and can reduce the impact of low-cited data during classification. However, classification results represent total citations within a certain period, simplifying citation volume processing [**?**] and thus cannot judge citation trend changes over time. Classification models lack unified classification standards, often custom-defined by researchers based on their datasets, with even the same researcher using different standards across studies, showing the coarse-grained nature of classification methods and limiting research applicability [**?**]. Some papers using machine learning for citation prediction are shown in Table 8 .

Analysis of Common Problems in Citation Frequency Prediction Research

Overall, whether defined as regression or classification, citation prediction research shares common problems.

Unclear causal relationship between influencing factors and prediction results

Research on influencing factors and citation frequency primarily examines correlation rather than causation. Correlation does not guarantee good predictive

performance in models. With numerous influencing factors, research has covered many aspects: paper/content-related, author-related, journal-related, and others including time, altmetric, and network features. However, different factors may produce different effects in different datasets. For example, in the KDDCUP dataset, J. Chen and C. Zhang found content features more important [**?**], while in the ArnetMiner dataset, X. Zhu and Z. Ban found author features more important [**?**].

Lack of diversity in research sample data

Citation prediction research samples are relatively homogeneous, mostly focusing on science, engineering, and medical literature. Although some studies compare disciplines, they do not cross the boundary between natural and social sciences, lacking comprehensiveness. The ArnetMiner and AMiner datasets are commonly used public datasets for computer science literature, with biomedical datasets also common but humanities and social science datasets very rare. Most datasets come from foreign databases. Existing research shows large differences between datasets from different fields, raising questions about whether these prediction methods remain applicable when migrated to domestic or humanities/social science datasets.

Unclear relationship between result applicability and prediction period

Prediction aims for practical application, but most papers do not explain what period is appropriate. Among many studies, prediction periods vary widely, with research goals of predicting short-term or long-term impact measured by citations in future periods of varying lengths (1, 5, 10, or even 31 years). Different researchers use different data, resulting in different periods. Few papers study the overall data to identify effective citation time windows. Overly long citation windows cause information lag and invalid predictions, while overly short windows may reduce model accuracy.

Weak interpretability of model evaluation

Citation prediction requires evaluation criteria such as coefficient of determination r^2, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Accuracy (ACC). However, many studies only provide evaluation values to judge model quality without detailed explanation. In fact, evaluation values are calculated from actual and predicted values. For example, MAE is mean absolute error, and when judging value magnitude, it should be compared with actual values to see the error range relative to real values, not just comparing error values across methods.

Suggestions for Improving Citation Prediction Research Quality

Addressing the common problems identified above, this paper offers suggestions to help researchers improve research quality.

Clarify problem preconditions and propose innovative prediction

paths

When solving practical problems, researchers often add preconditions to simplify complexity. What problems emerge when removing these preconditions, and whether methods can be reproduced in practice, warrant consideration. In research using dynamic heterogeneous information networks for newly published papers, Jiang et al. [**?**] argued that previous citation prediction relied on citations observed in the first few years after publication (leading citation values) to predict long-term citations. However, many papers reach peak citation impact in the first few years, failing to demonstrate leading value. In fast-updating fields like machine learning, waiting 3-5 years to predict impact is unrealistic. They therefore challenged: generating citation time series for new papers without any leading value, solving the "cold start" problem in time series tasks. They proposed an end-to-end framework from heterogeneous information network to time series, predicting single-paper citation frequency. The core idea is a transformation: learning from heterogeneous networks composed of keywords, authors, venues, and papers to estimate a pseudo-leading value and map it to future citation time series, converting network information into time series information.

Improve influencing factor extraction methods

The above review comprehensively discussed influencing factors that may become model features. However, making these features better express needed information requires innovation and application from micro-level operational perspectives. In research extracting advanced semantic features to learn citation time series [**?**], the core was obtaining semantic information from metadata text, encoding sentences using Doc2Vec, then extracting advanced (paragraph-level) semantic features through Bi-LSTM and attention mechanisms, finally learning citation prediction tasks by integrating early citations. This proved metadata semantic features useful for improving citation prediction performance, offering a promising method.

Topic-related feature research also mines text content (titles, abstracts), but differs in feature granularity. Topic features describe entire documents, commonly extracted using LDA and its improved models to form a fixed number of topics through parameter adjustment in corpora, with relatively coarse granularity where some papers may not find appropriate topics. Metadata semantic features, based on Doc2Vec, further use Bi-LSTM and attention mechanisms for semantic mining with finer granularity, enabling each paper to find its specific semantic features.

Expand research samples

Most citation prediction studies use single datasets, so results are not universally applicable to other datasets. Research shows large differences between field datasets. To make results more generalizable, more comprehensive datasets should be used for comparative studies across significantly different fields, analyzing reasons for different prediction results to make research more rigorous and comprehensive. In G. Abramo et al. [**?**], 123,128 Italian publications from

WoS were studied, finding different disciplines had different model applicability. All literature was classified into 12 subject categories: "Biology," "Biomedical Research," "Chemistry," "Clinical Medicine," "Earth and Space Sciences," "Economics," "Engineering," "Law, Politics and Sociology," "Mathematics," "Multidisciplinary," "Physics," and "Psychology." Results showed "Economics" had the largest weight for early citations in both models, while "Psychology" was opposite; life science fields had different average early citation weight coefficients; "Law, Politics and Sociology," "Engineering," and "Multidisciplinary" showed obvious early impact.

Apply mathematical modeling thinking to improve model interpretability

The aforementioned empiricism-based parameter-tuning machine learning and deep learning modeling methods lack mathematical tools to diagnose and evaluate neural network feature expression capabilities, resulting in weak interpretability. In the modeling process, appropriate methods can be sought based on research needs. Mathematical modeling thinking analyzes problems from a mathematical perspective in real contexts, proposing questions, building models, determining parameters, solving models, and ultimately solving practical problems. The following modeling methods fully demonstrate mathematical thinking and use mathematical tools for quantitative explanation, fully showing model interpretability.

In research on paper citation dynamics mechanisms, M. Wang et al. [**?**] started from the question "Can paper citation patterns predict long-term impact?" First, they identified three basic mechanisms driving citations: highly cited papers are more likely to be cited again; papers have aging effects where novelty eventually disappears; and papers have intrinsic differences. Combining these factors, they derived a probability model for paper citations: ( ) ~ , where i explains intrinsic differences. Since intrinsic differences like novelty and importance depend on multiple intangible and subjective dimensions, the study avoided assessing a paper' s intrinsic value, treating appropriate i as a comprehensive measure of a paper' s intrinsic differences within the total sample; Cit is citations paper i receives at time t after publication; di(t) is the decay rate of paper i at time t after publication. Total cumulative citations can be solved through calculus.

The innovation lies in treating citation prediction as a continuous probability problem, deriving probability density functions to obtain distributions and thus future citations. Compared to machine learning and deep learning methods with similar accuracy, this modeling approach offers stronger interpretability.

Conclusion and Outlook

In summary, in the era of big data and artificial intelligence, citation prediction research content continuously updates, generating new influencing factor indicators and prediction methods. This paper systematically reviewed from "influencing factors" to "research objects" and "research methods," summarizing

current problems and proposing corresponding suggestions based on previous research.

Future research should deepen theoretical studies, strengthen the rational use of influencing factor indicators and research methods, identify appropriate research cycles, establish unified evaluation systems, and improve theoretical foundations. On this basis, efforts should focus on solving practical problems, fully applying macro-level mathematical modeling thinking and implementing micro-level specific operational methods. Using transformation thinking, complex practical problems can be converted into multiple simple problems solved one by one, enabling models to be fully applied in practical problems.

References

[?] BARABASI A L, SONG C, WANG D. Publishing: a handful of papers dominates citation. Nature, 2012, 491(7422): 40.

[?] DIDEGAH F, THELWALL M. Determinants of research citation impact in nanoscience and nanotechnology. Journal of the American Society for Information Science and Technology, 2013, 64(5): 1055-1064.

[?] BUELA-CASAL G, ZYCH I. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals. Psicothema, 2010, 22(2): 270-275.

[?] FU L D, ALIFERIS C F. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics, 2010, 85(1): 257-270.

[?] YAN Y, TIAN S, ZHANG J. The impact of a paper' s new combinations and new components on its citation. Scientometrics, 2019, 122(2): 895-913.

[?] CHAKRABORTY T, KUMAR S, GOYAL P, et al. Towards a stratified learning approach to predict future citation counts. IEEE/ACM joint conference on digital libraries (Jcdl): IEEE, 2014: 351-360.

[?] 柴嘉琪, 陈仕吉. 论文新颖性测度研究综述. 农业图书情报学报, 2020, 32(10): 56-61.

[?] ANTONIOU G A, ANTONIOU S A, GEORGAKARAKOS E I, et al. Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature. Annals of vascular surgery, 2015, 29(2): 286-292.

[?] 魏瑞斌. 论文平均引用时差与被引频次相关性分析. 情报杂志, 2018, 37(2): 135-141.

[?] ROTH C, WU J, LOZANO S. Assessing impact and quality from local dynamics of citation networks. Journal of informetrics, 2013, 6(1): 111-120.

[?] BARNETT G A, FINK E L. Impact of the internet and scholar age distribution on academic citation age. Journal of the American Society for Information Science and Technology, 2008, 59(4): 526-534.

[?] BORNMANN L, SCHIER H, MARX W, et al. What factors determine citation counts of publications in chemistry besides their quality? Journal of informetrics, 2012, 6(1): 11-18.

[?] BISCARO C, GIUPPONI C. Co-authorship and bibliographic coupling network effects on citations. Plos one, 2014, 9(6): e99502.

[?] LEIMU R, KORICHEVA J. What determines the citation frequency of ecological papers? Trends in Ecology & Evolution, 2005, 20(1): 28-32.

[?] JAMALI H R, NIKZAD M. Article title type and its relation with the number of downloads and citations. Scientometrics, 2011, 88(2): 653-661.

[?] ROSTAMI F, MOHAMMADPOORASL A, HAJIZADEH M. The effect of characteristics of title on citation rates of articles. Scientometrics, 2014, 98(3): 2007-2010.

[?] MCCABE M J, SNYDER C M. Does online availability increase citations? theory and evidence from a panel of economics and business journals. Review of economics and statistics, 2015, 97(1): 144-165.

[?] STREMERSCH S, CAMACHO N, VANNESTE S, et al. Unraveling scientific impact: citation types in marketing journals. International journal of research in marketing, 2015, 32(1): 64-77.

[?] ZHANG X, XIE Q, SONG M. Measuring the impact of novelty, bibliometric, and academic-network factors on citation count using a neural network. Journal of informetrics, 2021, 15(2): 101-140.

[?] MONTEFUSCO A M, NASCIMENTO F P, SENNES L U, et al. Influence of international authorship on citations in Brazilian medical journals: a bibliometric analysis. Scientometrics, 2019, 119(3): 1487-1496.

[?] 魏瑞斌. 论文平均引用时差与被引频次相关性分析. 情报杂志, 2018, 37(2): 135-141.

[?] BORNMANN L, DANIEL H-D. Selecting manuscripts for a high-impact journal through peer review: a citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. JASIST, 2008, 59(12): 1841-1852.

[?] SKILTON P F. Does the human capital of teams of natural science authors predict citation frequency? Scientometrics, 2009, 78(3): 525-542.

[?] GUILERA G, GóMEZ-BENITO J, HIDALGO M D. Citation analysis in research on differential item functioning. Quality & quantity, 2009, 44(6): 1249-1258.

[?] AKSNES D W. Characteristics of highly cited papers. Research evaluation, 2003, 12(3): 159-170.

[?] ONODERA N, YOSHIKANE F. Factors affecting citation rates of research articles. Journal of the Association for Information Science and Technology, 2015, 66(4): 739-764.

[?] COLLET F, ROBERTSON D A, LUP D. When does brokerage matter? citation impact of research teams in an emerging academic field. Strategic organization, 2014, 12(3): 157-179.

[?] YU T, YU G, LI P Y, et al. Citation impact prediction for scientific papers using stepwise regression analysis. Scientometrics, 2014, 101(2): 1233-1252.

[?] AIN Q-U, RIAZ H, AFZAL M T. Evaluation of h-index and its citation intensity based variants in the field of mathematics. Scientometrics, 2019, 119(1): 187-211.

[?] AMARA N, LANDRY R, HALILEM N. What can university administrators do to increase the publication and citation scores of their faculty members? Scientometrics, 2015, 103(2): 489-530.

[?] NOSEK B A, GRAHAM J, LINDNER N M, et al. Cumulative and career-stage citation impact of social-personality psychology programs and their members. Personality & social psychology bulletin, 2010, 36(10): 1273-1289.

[?] BORSUK R M, BUDDEN A E, LEIMU R, et al. The influence of author gender, national language and number of authors on citation rate in Ecology. Open ecology journal, 2009, 2(1): 25-28.

[?] PENG T-Q, ZHU J J H. Where you publish matters most: a multilevel analysis of factors affecting citations of internet studies. Journal of the American Society for Information Science and Technology, 2012, 63(9): 1789-1801.

[?] VAN DER POL C B, MCINNES M D, PETRCICH W, et al. Is quality and completeness of reporting of systematic reviews and meta-analyses published in high impact radiology journals associated with citation rates? Plos one, 2015, 10(3): e0119892.

[?] ROLDAN-VALADEZ E, RIOS C. Alternative bibliometrics from impact factor improved the esteem of a journal in a 2-year-ahead annual-citation calculation: multivariate analysis of gastroenterology and hepatology journals. European journal of gastroenterology & hepatology, 2015, 27(2): 115-122.

[?] ZHU X P, BAN Z J. Citation count prediction based on academic network features. Proceedings 2018 IEEE 32nd international conference on advanced information networking and applications (Aina). New York: IEEE, 2018: 534-541.

[?] DING Y, JACOB E K, ZHANG Z X, et al. Perspectives on social tagging. Journal of the American Society for Information Science and Technology, 2009, 60(12): 2388-2401.

[?] 孔玲, 王效岳, 于纯良, 等. 学术论文离被引有多远——基于影响因素与预测方法的文献述评. 情报资料工作, 2019, 40(6): 63-72.

[?] YAN R, HUANG C, TANG J, et al. To better stand on the shoulder of giants. Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries. New York: ACM, 2012: 51-60.

[?] BUTUN E, KAYA M. Predicting citation count of scientists as a link prediction problem. IEEE transactions on cybernetics, 2020, 50(10): 4518-4529.

[?] 耿骞, 景然, 靳健, 等. 学术论文引用预测及影响因素分析. 图书情报工作, 2018, 62(14): 29-40.

[?] RUAN X M, ZHU Y Y, LI J, et al. Predicting the citation counts of individual papers via a BP neural network. Journal of informetrics, 2020, 14(3): 101039.

[?] LOKKER C, MCKIBBON K A, MCKINLAY R J, et al. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. BMJ, 2008, 336(7645): 655-657.

[?] ABRAMO G, D' ANGELO C A, FELICI G. Predicting publication long-term impact through a combination of early citations and journal impact factor. Journal of informetrics, 2019, 13(1): 32-49.

[?] BORNMANN L, LEYDESDORFF L, WANG J. How to improve the prediction based on citation impact percentiles for years shortly after the publication date? Journal of informetrics, 2014, 8(1): 175-180.

[?] 程子轩, 张向先, 郭顺利. 基于作者特征和期刊特征的学术论文被引频次预测模型构建与分析. 情报科学, 2021, 39(3): 179-184, 192.

[?] YAN R, TANG J, LIU X, et al. Citation count prediction: learning to estimate future citations for literature. Proceedings of the 20th ACM international conference on information and knowledge management. Glasgow, Scotland: Association for Computing Machinery, 2011: 1247-1252.

[?] CHEN J P, ZHANG C X. Predicting citation counts of papers. Proceedings of 2015 IEEE 14th international conference on cognitive informatics & cognitive computing. New York: IEEE, 2015: 434-440.

[?] AFZAL M, PARK B J, HUSSAIN M, et al. Deep learning based biomedical literature classification using criteria of scientific rigor. Electronics, 2020, 9(8): 9081253.

[?] ABRISHAMI A, ALIAKBARY S. Predicting citation counts based on deep neural network learning techniques. Journal of informetrics, 2019, 13(2): 485-499.

[?] YUAN S, TANG J, ZHANG Y, et al. Modeling and predicting citation count via recurrent neural network with long short-term memory. arXiv preprint arXiv:1811.02129.

[?] WEN J Q, WU L Y, CHAI J P. Paper citation count prediction based on recurrent neural network with gated recurrent unit. Proceedings of 2020 IEEE 10th international conference on electronics information and emergency communication. New York: IEEE, 2020: 303-306.

[?] XU J, LI M, JIANG J, et al. Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network. IEEE access, 2019, 7: 63396-63406.

[?] DONG Y, JOHNSON R A, CHAWLA N V. Will this paper increase your h-index? Proceedings of the eighth ACM international conference on Web search and data mining. New York: ACM, 2015: 149-158.

[?] IBANEZ A, LARRANAGA P, BIELZA C. Predicting citation count of Bioinformatics papers within four years of publication. Bioinformatics, 2009, 25(24): 3303-3309.

[?] WANG M, YU G, YU D. Mining typical features for highly cited papers. Scientometrics, 2011, 87(3): 695-706.

[?] MA A, LIU Y, XU X, et al. A deep-learning based citation count prediction model with paper metadata semantic features. Scientometrics, 2021, 126: 6803-6823.

[?] JIANG S, KOCH B, SUN Y. HINTS: citation time series prediction for new publications via dynamic heterogeneous information network embedding. Proceedings of the Web conference 2021. New York: ACM, 2021: 3158-3167.

Author Contributions:
Zhang Sufang: Framework guidance, proposed revisions, paper proofreading and finalization.
Liu Huimin: Paper writing, data compilation.

Note: Figure translations are in progress. See original paper for figures.